

Scim: Intelligent Skimming Support for Scientific Papers

Raymond Fok (presenter), Hita Kambhamettu, Luca Soldaini, Jonathan Bragg,
Kyle Lo, Marti A. Hearst, Andrew Head, Daniel S. Weld

IUI 2023



The Explosion of Scientific Literature



Rate of academic publishing increased
~4% per year, over the past decade.

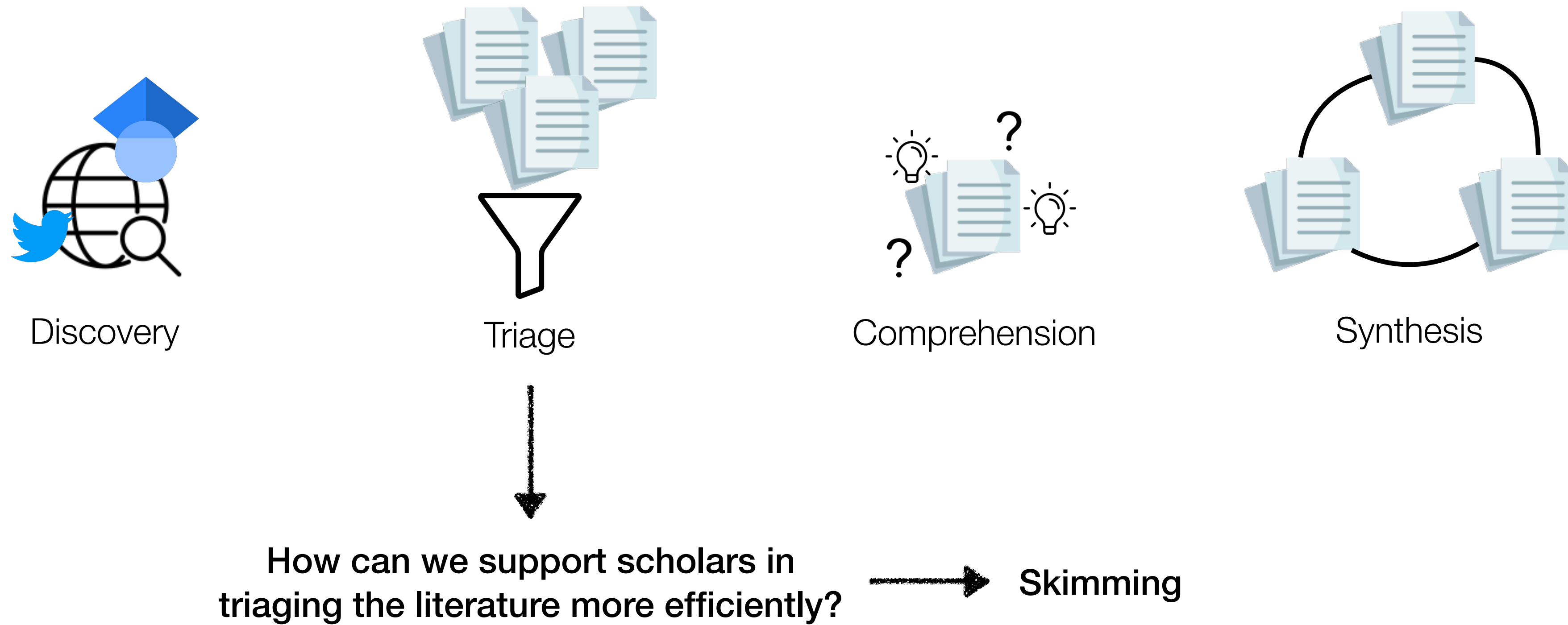


15k+ submissions per month



200M+ papers
across all of science

Consuming Scientific Literature



Augmenting Static PDFs

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **Bidirectional Encoder Representations from Transformers**. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction


Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **Bidirectional Encoder Representations from Transformers**. BERT alleviates the previously mentioned unidirectionality constraint by using a “masked language model” (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

Augmentation through automatic highlighting



BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Abstract

We introduce a new language representation model called **BERT**, which stands for **Bidirectional Encoder Representations from Transformers**. Unlike recent language representation models (Peters et al., 2018a; Radford et al., 2018), BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. As a result, the pre-trained BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for a wide range of tasks, such as question answering and language inference, without substantial task-specific architecture modifications.

BERT is conceptually simple and empirically powerful. It obtains new state-of-the-art results on eleven natural language processing tasks, including pushing the GLUE score to 80.5% (7.7% point absolute improvement), MultiNLI accuracy to 86.7% (4.6% absolute improvement), SQuAD v1.1 question answering Test F1 to 93.2 (1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (5.1 point absolute improvement).

1 Introduction

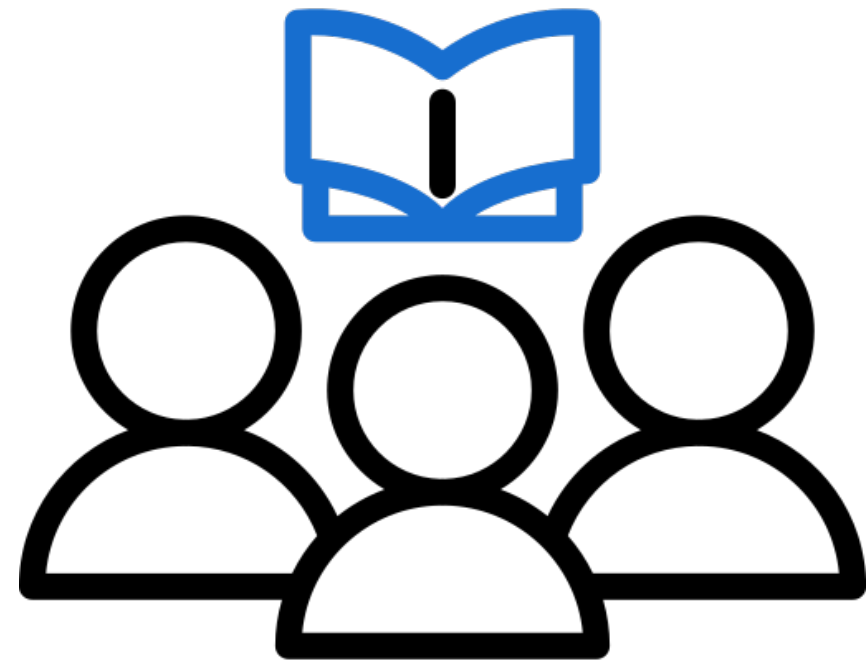
Language model pre-training has been shown to be effective for improving many natural language processing tasks (Dai and Le, 2015; Peters et al., 2018a; Radford et al., 2018; Howard and Ruder, 2018). These include sentence-level tasks such as natural language inference (Bowman et al., 2015; Williams et al., 2018) and paraphrasing (Dolan and Brockett, 2005), which aim to predict the relationships between sentences by analyzing them holistically, as well as token-level tasks such as named entity recognition and question answering, where models are required to produce fine-grained output at the token level (Tjong Kim Sang and De Meulder, 2003; Rajpurkar et al., 2016).

There are two existing strategies for applying pre-trained language representations to downstream tasks: *feature-based* and *fine-tuning*. The feature-based approach, such as ELMo (Peters et al., 2018a), uses task-specific architectures that include the pre-trained representations as additional features. The fine-tuning approach, such as the Generative Pre-trained Transformer (OpenAI GPT) (Radford et al., 2018), introduces minimal task-specific parameters, and is trained on the downstream tasks by simply fine-tuning *all* pre-trained parameters. The two approaches share the same objective function during pre-training, where they use unidirectional language models to learn general language representations.

We argue that current techniques restrict the power of the pre-trained representations, especially for the fine-tuning approaches. The major limitation is that standard language models are unidirectional, and this limits the choice of architectures that can be used during pre-training. For example, in OpenAI GPT, the authors use a left-to-right architecture, where every token can only attend to previous tokens in the self-attention layers of the Transformer (Vaswani et al., 2017). Such restrictions are sub-optimal for sentence-level tasks, and could be very harmful when applying fine-tuning based approaches to token-level tasks such as question answering, where it is crucial to incorporate context from both directions.

In this paper, we improve the fine-tuning based approaches by proposing BERT: **Bidirectional Encoder Representations from Transformers**. BERT alleviates the previously mentioned unidirectionality constraint by using a “masked language model” (MLM) pre-training objective, inspired by the Cloze task (Taylor, 1953). The masked language model randomly masks some of the tokens from the input, and the objective is to predict the original vocabulary id of the masked

What do scholars want from a skimming aid?



Formative user study with 10 scholars

Design Motivations

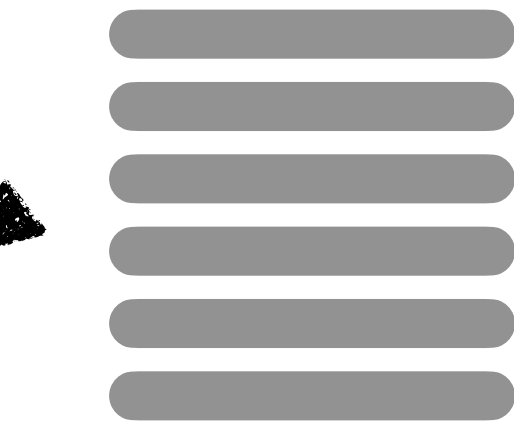
An intelligent highlighting-based skimming aid should...

- Augment readers' skimming practices.
- Highlight diverse kinds of content corresponding to information needs.
- Support skimming through text-dense sections.
- Provide user customization

The Approach



Paper
(as PDF)



Sentences
(with metadata)

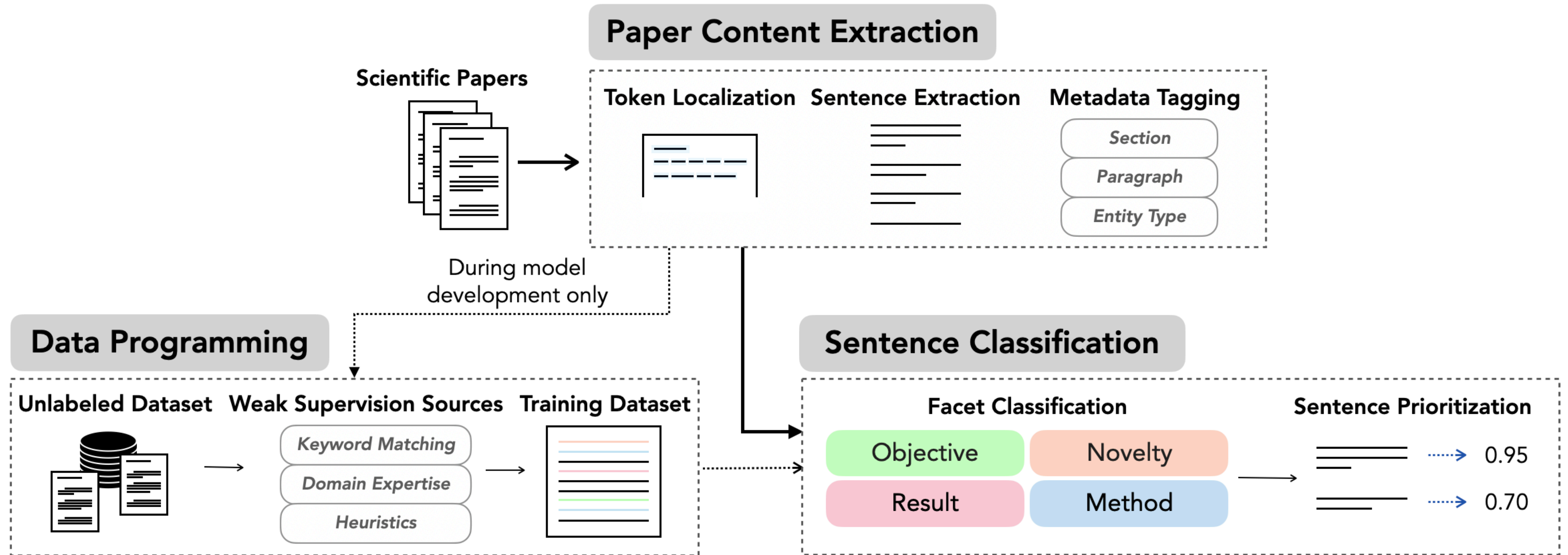


Important sentences.



Skimming UI

The Paper Processing Pipeline



Aligning Highlights to Information Needs

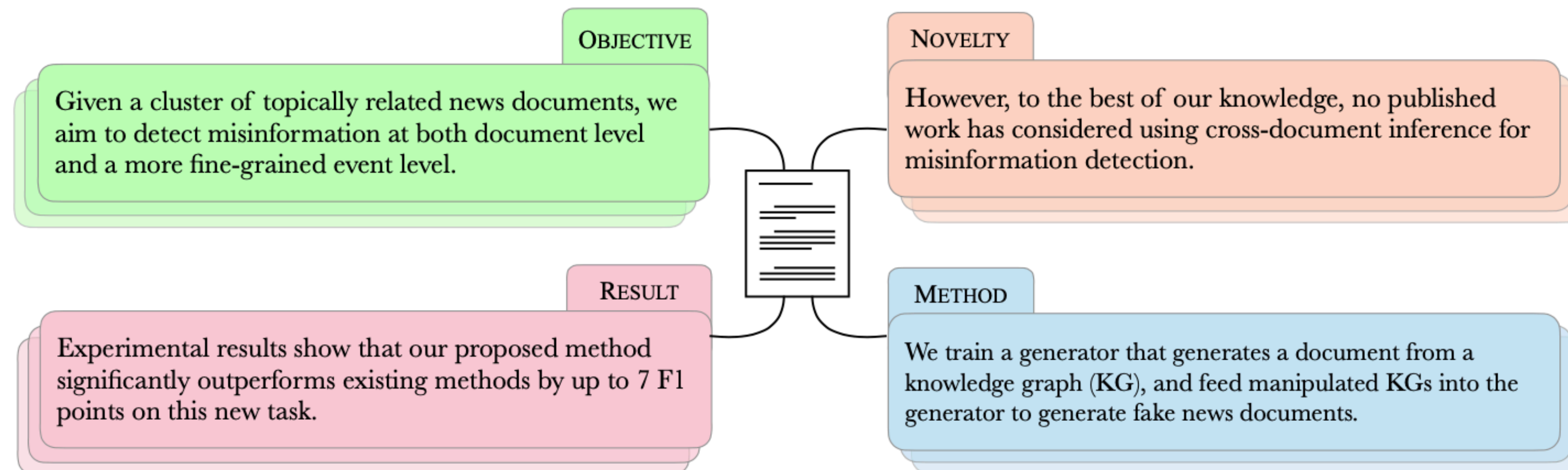
Readers often share some common information needs when skimming.

What are their key results?

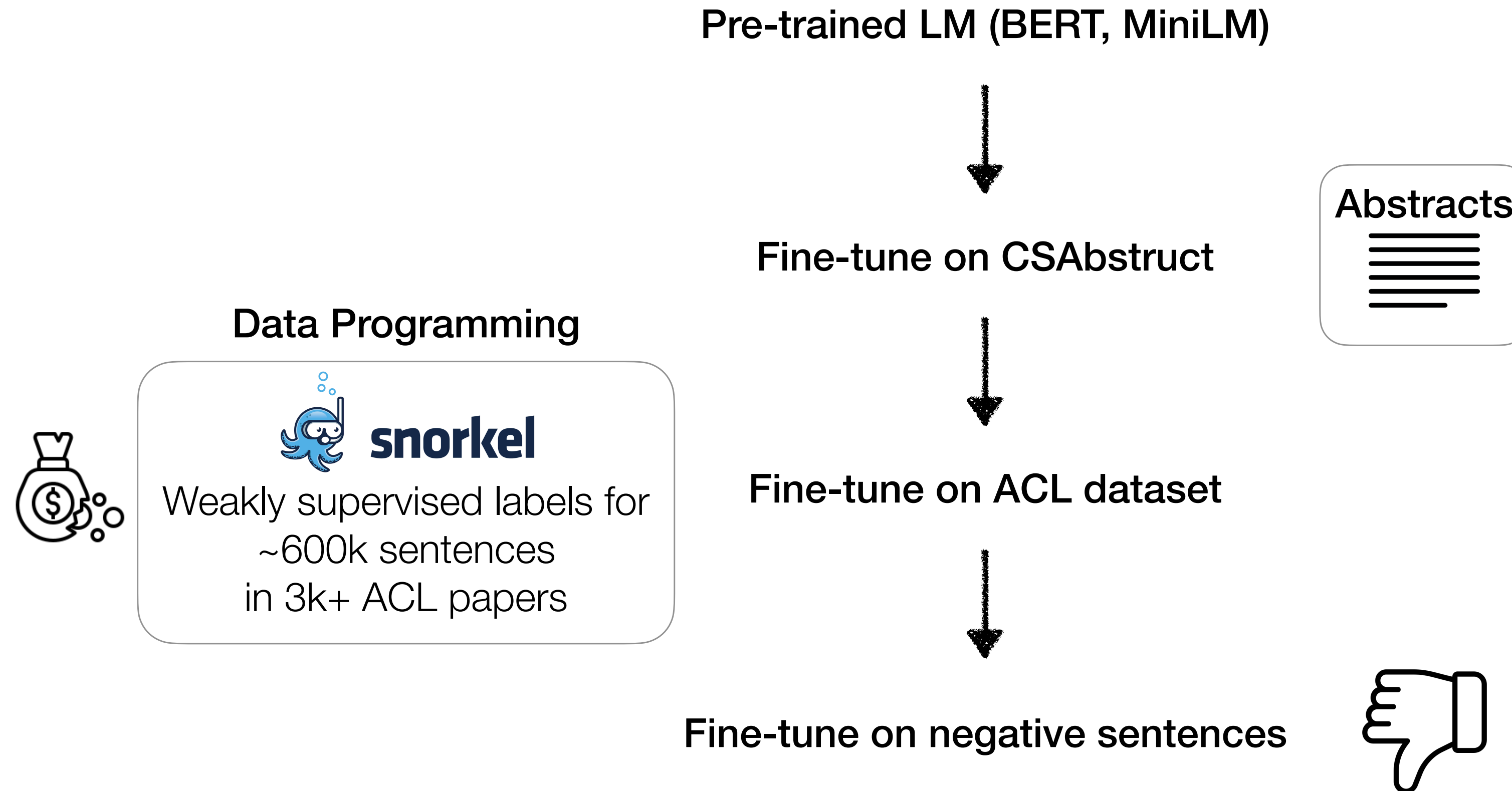


What's new about this work?

Scim identifies **important** sentences, and then labels them with common **information facets**: Objective, Novelty, Result, and Method.



Classifying Faceted Information



An Augmented Skimming Interface

107 (1 of 9) Objective Novelty Method Result Disable skimming

tention scores between tokens in self-attention mechanism, are sometimes ineffective as they are learned implicitly without the guidance of explicit semantic knowledge. Thus, we aim to infuse explicit external knowledge into pre-trained language models to further boost their performance. Existing works of knowledge infusion largely depend on multi-task learning frameworks, which are inefficient and require large-scale re-training when new knowledge is considered. In this paper, we propose a novel and generic solution, KAM-BERT, which directly incorporates knowledge-generated attention maps into the self-attention mechanism. It requires only a few extra parameters and supports efficient fine-tuning once new knowledge is added. KAM-BERT achieves consistent improvements on various academic datasets for natural language understanding. It also outperforms other state-of-the-art methods which conduct knowledge infusion into transformer-based architectures. Moreover, we apply our model to an industry-scale ad relevance application and show its advantages in the real-world scenario.

1 Introduction

Language models pre-trained by a large text corpus have shown superior performances on a wide range of natural language processing tasks. Many advanced models based on the transformer architectures achieve state-of-the-art results on various NLP benchmarks. Existing literature (Jawahar et al., 2019; Hewitt and Manning, 2019) shows that pre-training enables a model to capture syntactic and semantic information in the self-attention mechanism. However, the attention maps, which

good chance to improve the quality of attention scores as well as the performance of downstream applications.

Recently, there have been multiple attempts for incorporating knowledge into transformer architectures. ERNIE (Zhang et al., 2019) and KEPLER (Wang et al., 2019) utilize both large-scale textual corpora and knowledge graphs to train a representation model in a multi-task learning framework. They need to be retrained from scratch when injecting new knowledge, which is inefficient and can not benefit from existing pre-trained checkpoints. K-Adapter (Wang et al., 2020) integrates additional neural models to capture different kinds of knowledge. It enables adaptation based on pre-

Highlight important information

In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above. First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism. In Figure 1, the attention map of a query-ad pair is visualized, and the goal is to judge if the query and ad text are semantically relevant. As shown in the figure, BERT misclassifies this pair as irrelevant, probably because it does not understand the query word “glipizide”, which rarely appears in the pre-training corpus. In fact, “glipizide” is a kind of medicine and highly related to the word “Pharmacy” in ad text, so this

Objective (2)

Novelty (13)

Method (38)

Result (19)

well as the performance of downstream applications.

In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above.

First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism.

To address the above motivation, we propose a novel architecture, namely KAM-BERT (Knowledge-assisted Attention Maps for BERT).

Specifically, we consider three kinds of semantic knowledge to guide the self-attention mechanism, i.e., entity, phrase

An Augmented Skimming Interface

The interface displays a document with text and a sidebar on the right. The sidebar contains a list of highlights categorized by Objective (2), Novelty (13), Method (38), and Result (19). A callout box points to the sidebar with the text "Summarize highlights in a sidebar view".

107 (1 of 9) Objective Novelty Method Result Disable skimming

tention scores between tokens in self-attention mechanism, are sometimes ineffective as they are learned implicitly without the guidance of explicit semantic knowledge. Thus, we aim to infuse explicit external knowledge into pre-trained language models to further boost their performance. Existing works of knowledge infusion largely depend on multi-task learning frameworks, which are inefficient and require large-scale re-training when new knowledge is considered. In this paper, we propose a novel and generic solution, KAM-BERT, which directly incorporates knowledge-generated attention maps into the self-attention mechanism. It requires only a few extra parameters and supports efficient fine-tuning once new knowledge is added. KAM-BERT achieves consistent improvements on various academic datasets for natural language understanding. It also outperforms other state-of-the-art methods which conduct knowledge infusion into transformer-based architectures. Moreover, we apply our model to an industry-scale ad relevance application and show its advantages in the real-world scenario.

1 Introduction

Language models pre-trained by a large text corpus have shown superior performances on a wide range of natural language processing tasks. Many advanced models based on the transformer architectures achieve state-of-the-art results on various NLP benchmarks. Existing literature (Jawahar et al., 2019; Hewitt and Manning, 2019) shows that pre-training enables a model to capture syntactic and semantic information in the self-attention mechanism. However, the attention maps, which

good chance to improve the quality of attention scores as well as the performance of downstream applications.

Recently, there have been multiple attempts for incorporating knowledge into transformer architectures. ERNIE (Zhang et al., 2019) and KEPLER (Wang et al., 2019) utilize both large-scale textual corpora and knowledge graphs to train a representation model in a multi-task learning framework. They need to be retrained from scratch when injecting new knowledge, which is inefficient and can not benefit from existing pre-trained checkpoints. K-Adapter (Wang et al., 2020) integrates additional neural models to capture different kinds of knowledge. It enables adaptation based on pre-trained language models. K-Adapter instructs the self-attention mechanism to introduce a relationship between the input and the original model.

In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above. First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism. In Figure 1, the attention map of a query-ad pair is visualized, and the goal is to judge if the query and ad text are semantically relevant. As shown in the figure, BERT misclassifies this pair as irrelevant, probably because it does not understand the query word “glipizide”, which rarely appears in the pre-training corpus. In fact, “glipizide” is a kind of medicine and highly related to the word “Pharmacy” in ad text, so this

well as the performance of downstream applications.

In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above.

First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism.

To address the above motivation, we propose a novel architecture, namely KAM-BERT (Knowledge-assisted Attention Maps for BERT).

Specifically, we consider three kinds of semantic knowledge to guide the self-attention mechanism, i.e., entity, phrase

Summarize highlights in a sidebar view

An Augmented Skimming Interface

107 (1 of 9) Objective Novelty Method Result Disable skimming

tention scores between tokens in self-attention mechanism, are sometimes ineffective as they are learned implicitly without the guidance of explicit semantic knowledge. Thus, we aim to infuse explicit external knowledge into pre-trained language models to further boost their performance. Existing works of knowledge infusion largely depend on multi-task learning frameworks, which are inefficient and require large-scale re-training when new knowledge is considered. In this paper, we propose a novel and generic solution, KAM-BERT, which directly incorporates knowledge-generated attention maps into the self-attention mechanism. It requires only a few extra parameters and supports efficient fine-tuning once new knowledge is added. KAM-BERT achieves consistent improvements on various academic datasets for natural language understanding. It also outperforms other state-of-the-art methods which conduct knowledge infusion into transformer-based architectures. Moreover, we apply our model to an industry-scale ad relevance application and show its advantages in the real-world scenario.

1 Introduction

Language models pre-trained by a large text corpus have shown superior performances on a wide range of natural language processing tasks. Many advanced models based on the transformer architectures achieve state-of-the-art results on various NLP benchmarks. Existing literature (Jawahar et al., 2019; Hewitt and Manning, 2019) shows that pre-training enables a model to capture syntactic and semantic information in the self-attention mechanism. However, the attention maps, which

good chance to improve the quality of attention scores as well as the performance of downstream applications.

Recently, there have been multiple attempts for incorporating knowledge into transformer architectures. ERNIE (Zhang et al., 2019) and KEPLER (Wang et al., 2019) utilize both large-scale textual corpora and knowledge graphs to train a representation model in a multi-task learning framework. They need to be retrained from scratch when injecting new knowledge, which is inefficient and can not benefit from existing pre-trained checkpoints. K-Adapter (Wang et al., 2020) integrates additional neural models to capture different kinds of knowledge. It enables adaptation based on pre-trained language models. However, it does not instruct the self-attention mechanism directly and introduces a relatively large number of parameters to the original model.

In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above. First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism. In Figure 1, the attention map of a query-ad pair is visualized, and the goal is to judge if the query and ad text are semantically relevant. As shown in the figure, BERT misclassifies this pair as irrelevant, probably because it does not understand the query word “glipizide”, which rarely appears in the pre-training corpus. In fact, “glipizide” is a kind of medicine and highly related to the word “Pharmacy” in ad text, so this

well as the performance of downstream applications.

In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above.

explicit knowledge into self-attention mechanism.

To address the above motivation, we propose a novel architecture, namely KAM-BERT (Knowledge-assisted Attention Maps for BERT).

Specifically, we consider three kinds of semantic knowledge to guide the self-attention mechanism, i.e., entity, phrase

Objective (2)
Novelty (13)
Method (38)
Result (19)

Show distribution of highlights at a glance

An Augmented Skimming Interface

107 (1 of 9) Objective Novelty Method Result Disable skimming

tention scores between tokens in self-attention mechanism, are sometimes ineffective as they are learned implicitly without the guidance of explicit semantic knowledge. Thus, we aim to infuse explicit external knowledge into pre-trained language models to further boost their performance. Existing works of knowledge infusion largely depend on multi-task learning frameworks, which are inefficient and require large-scale re-training when new knowledge is considered. In this paper, we propose a novel and generic solution, KAM-BERT, which directly incorporates knowledge-generated attention maps into the self-attention mechanism. It requires only a few extra parameters and supports efficient fine-tuning once new knowledge is added. KAM-BERT achieves consistent improvements on various academic datasets for natural language understanding. It also outperforms other state-of-the-art methods which conduct knowledge infusion into transformer-based architectures. Moreover, we apply our model to an industry-scale ad relevance application and show its advantages in the real-world scenario.

1 Introduction

Language models pre-trained by a large text corpus have shown superior performances on a wide range of natural language processing tasks. Many advanced models based on the transformer architectures achieve state-of-the-art results on various NLP benchmarks. Existing literature (Jawahar et al., 2019; Hewitt and Manning, 2019) shows that pre-training enables a model to capture syntactic and semantic information in the self-attention mechanism. However, the attention maps, which

good chance to improve the quality of attention scores as well as the performance of downstream applications.

Recently, there have been multiple attempts for incorporating knowledge into transformer architectures. ERNIE (Zhang et al., 2019) and KEPLER (Wang et al., 2019) utilize both large-scale textual corpora and knowledge graphs to train a representation model in a multi-task learning framework. They need to be retrained from scratch when injecting new knowledge, which is inefficient and can not benefit from existing pre-trained checkpoints. K-Adapter (Wang et al., 2020) integrates additional neural models to capture different kinds of knowledge. It enables adaptation based on pre-trained language models. However, it does not instruct the self-attention mechanism directly and introduces a relatively large number of parameters to the original model.

In this paper, we propose a novel and generic self-attention mechanism enhanced by explicit knowledge to address problems mentioned above. First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism. In Figure 1, the attention map of a query-ad pair is visualized, and the goal is to judge if the query and ad text are semantically relevant. As shown in the figure, BERT misclassifies this pair as irrelevant, probably because it does not understand the query word “glipizide”, which rarely appears in the pre-training corpus. In fact, “glipizide” is a kind of medicine and highly related to the word “Pharmacy” in ad text, so this

well as the performance of downstream applications.

In this paper, we propose a novel and

First, we show a failure case of query-ad matching, which motivates us to inject explicit knowledge into self-attention mechanism.

To address the above motivation, we propose a novel architecture, namely KAM-BERT (Knowledge-assisted Attention Maps for BERT).

Specifically, we consider three kinds of semantic knowledge to guide the self-attention mechanism, i.e., entity, phrase

Objective (2)

Novelty (13)

Method (38)

Result (19)

Support customization of highlights

Study 1: Usability Study



Participants

19 scholars (11 PhD, 5 Master's, 2 SWE, 1 industry researcher)



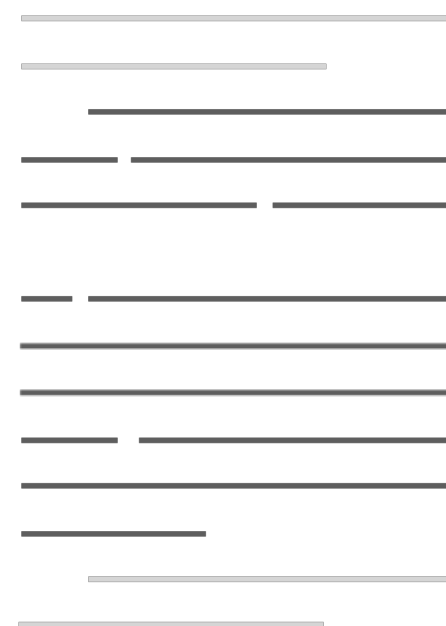
Papers

How does Scim impact readers' ability to search for specific kinds of information in a paper?
4 recent NLP papers



Interface Conditions

Standard Document Reader

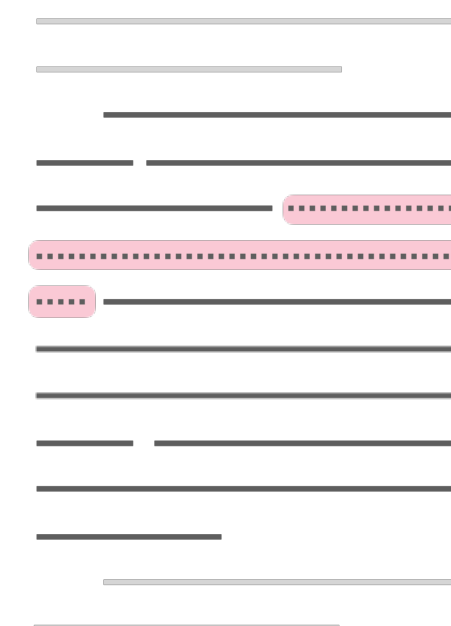


Scim

(answer in highlights)

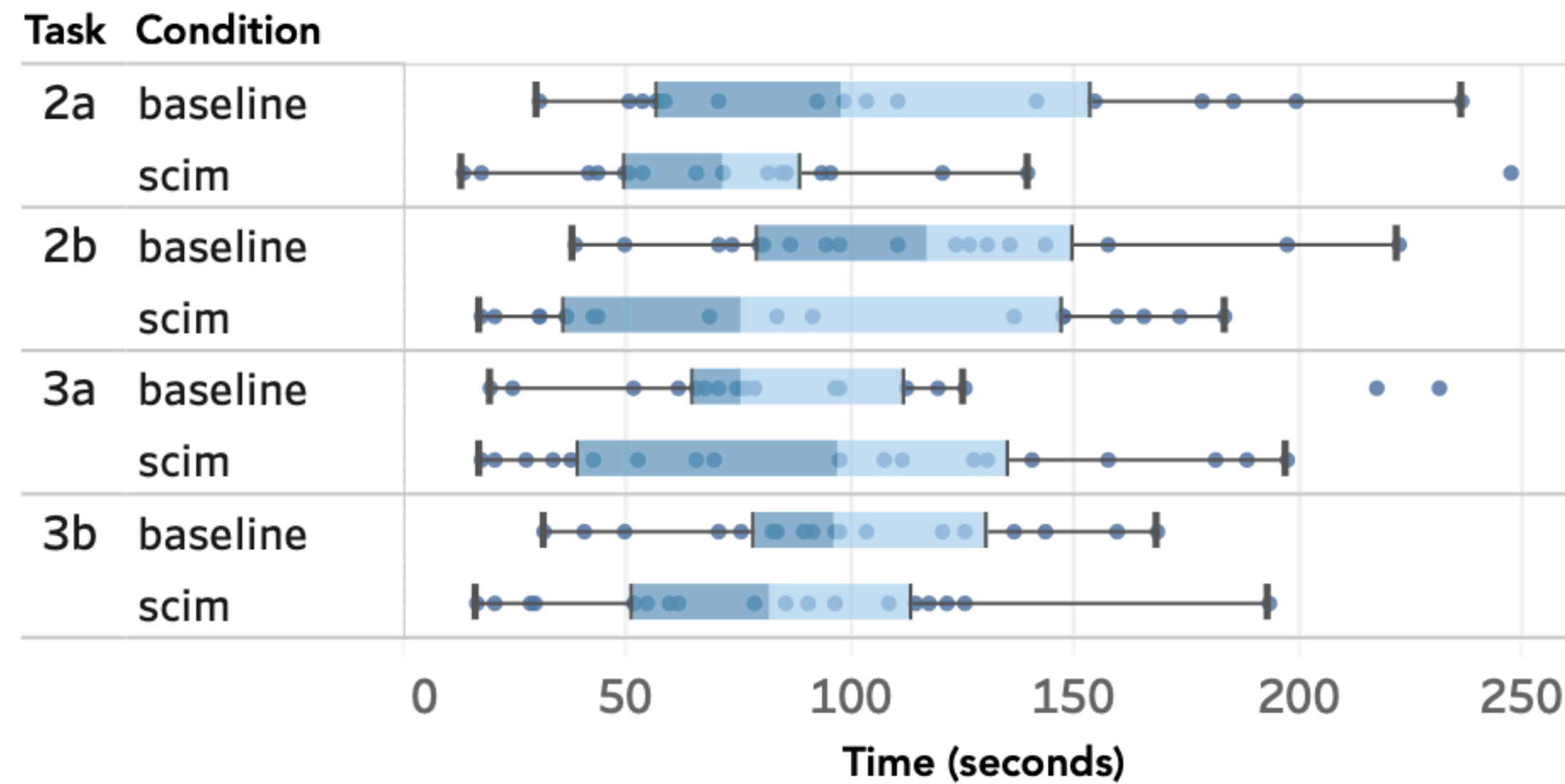


(answer not in highlights)

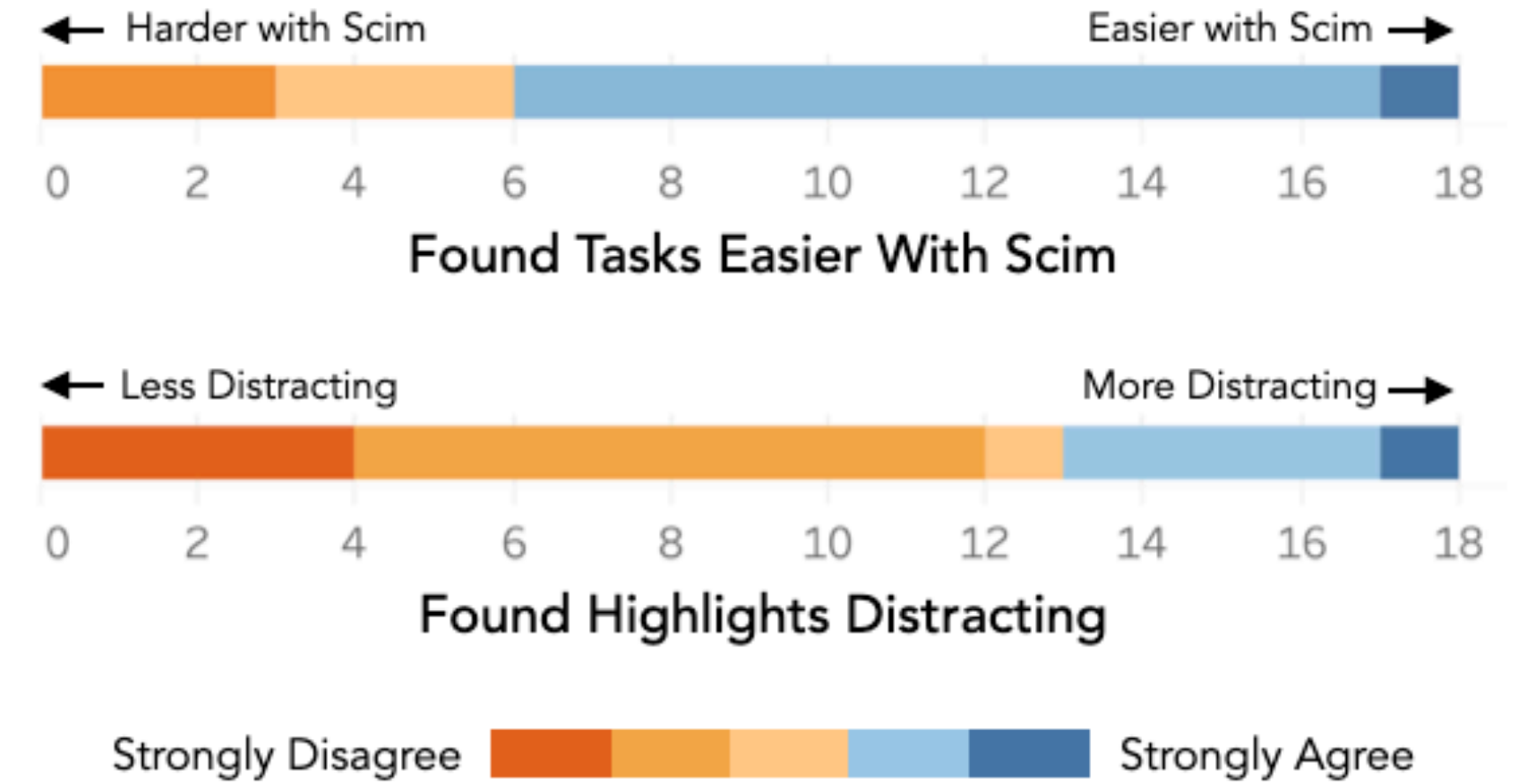


Usability Study - Key Findings

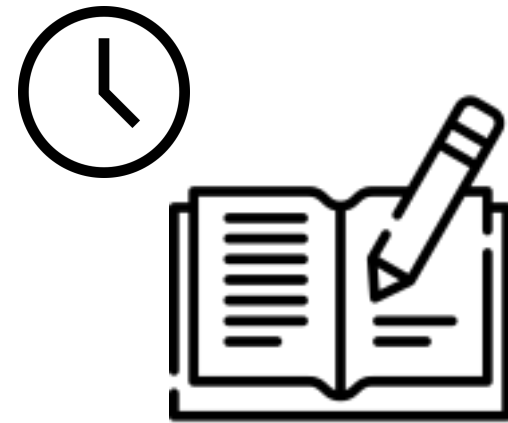
Overall, participants completed information-seeking tasks more quickly with Scim than with a standard document reader.



Participants felt information-seeking tasks were easier to complete with Scim, and highlights were not overly distracting.



Study 2: Diary Study



We conducted a **2-week long diary study**
with **12 scholars** (11 PhD, 1 Master's students)

skimming papers of their choice from the **proceedings of NAACL '22**

How would scholars use Scim for more realistic skimming tasks?

Participants skimmed **at least one paper per day**, and created a **diary entry**.

Diary entry prompts:

1. Did highlights help you skim this paper? Explain.
2. List one or more ways the system could have helped you better skim this paper.

Diary Study - Key Findings



Scim's highlights provided an [approachable, low-effort summary](#) of [high-value information](#) in a paper. Readers could skim only the highlights, yet easily gain more [context on-demand](#).



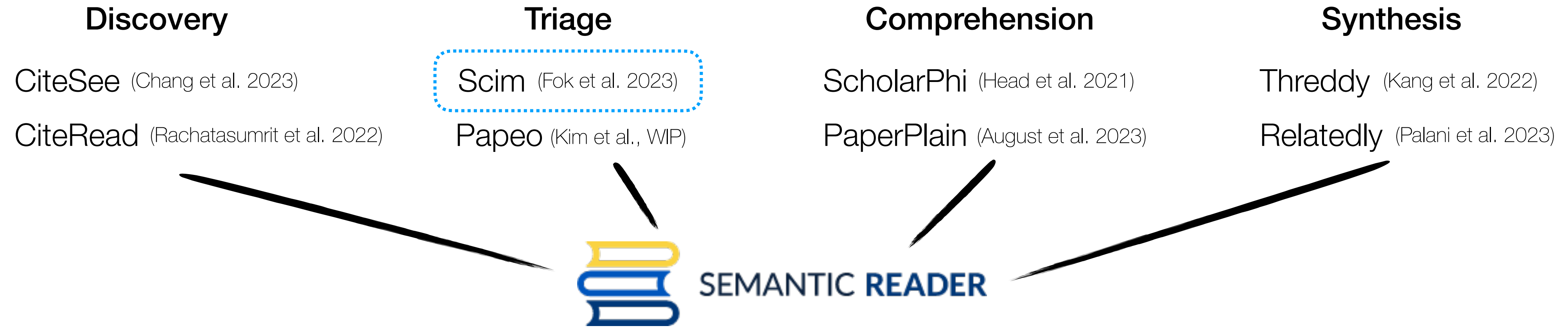
Participants noted Scim was particularly useful for:

- Skimming [dense](#) texts
- Skimming papers from [unfamiliar domains](#)
- Skimming with [low engagement](#)



Over the two weeks, participants noted learning how to better skim with an [imperfect AI assistant](#), i.e., with highlights that were potentially inaccurate or irrelevant.

AI-Augmented Scientific Reading and Writing



What's next?

How should we leverage LLMs (e.g., GPT-4) to support AI-augmented reading?

How should AI-powered assistants support the creation of academic literature (i.e., writing?)

Summary

Scim is an **augmented** reading interface leveraging **intelligent, faceted highlights** to support **skimming** of scientific papers.

Our findings suggest the desire for and promise of AI-powered skimming aids, yet also raise design challenges in **augmenting human cognition** without introducing undesirable **cognitive overload**.



Raymond Fok



Hita Kambhamettu



Luca Soldaini



Jonathan Bragg



Kyle Lo



Marti Hearst



Andrew Head



Dan Weld



Paper



Dataset



Code